基于 SEER 数据库应用贝叶斯网络构建亚洲肿瘤患者预后模型*

——以非小细胞肺癌为例

尹玢璨[1]　辛世超[1]　张　晗[1]　赵玉虹[1,2]

[1](中国医科大学医学信息学院　沈阳 110122)

[2](中国医科大学附属盛京医院　沈阳 110004)

摘要:【目的】利用 SEER 数据库, 找出对非小细胞肺癌患者预后生存的影响因素并预测患者预后生存状态, 指导肿瘤预后评价。【方法】采用单因素统计学方法及 Logistic 回归分析初步筛选预后相关因素, 利用贝叶斯网络方法构建患者术后生存预测模型, 并与其他三种常见的机器学习分类算法所建模型效能做比较。【结果】最终纳入模型的预后变量共 5 项, 包括年龄、肿瘤大小、组织学分级、肿瘤分期和受累淋巴结比率。贝叶斯网络所建模型对非小细胞肺癌患者生存状况预测准确率达到 72.87%。【局限】SEER 数据库内纳入的预后因素有限, 一定程度影响预测效果。【结论】贝叶斯网络可探寻变量间的关系并构建肺癌患者最优预后模型, 辅助医生判断患者预后情况及治疗效果, 优于决策树、支持向量机及人工神经网络三种模式。

关键词: 贝叶斯网络　非小细胞肺癌　预后　机器学习

分类号: R730.7　G35

1　引　言

肺癌是肿瘤患者死亡的主要原因, 其中非小细胞肺癌(Non-Small Cell Lung Cancer, NSCLC)约占所有肺癌病例的 83%, 其发病率为 40.60/10 万, 5 年生存率仅为 22.1%[1]。非小细胞肺癌发病率高且预后差, 对其预后的判断就尤为重要。目前临床医生通常根据手术病理分期判断预后, 但该分期仅考虑到肿瘤原发灶、区域淋巴结受累和远处转移三方面, 忽略了其他预后影响因素的作用, 预测效果差[2]。目前少有的预后研究多以单独或较少几个医疗机构为主要研究单位, 随访数据缺失多、数据量小、可信度差。临床上亟需有基于较大量数据、可信度高、预测效果好的非小细胞肺癌患者预后预测评估体系。

美国国家癌症研究所(National Cancer Institute, NCI)于 1973 年建立了监测、流行病学及预后数据库(Surveillance, Epidemiology and End Results, SEER), 是世界公认的肿瘤患者随访数据权威来源之一, 为临床研究提供了可靠的数据支持, 有学者利用此数据库, 采用简单统计学方法建立了横纹肌肉瘤等疾病生存预测模型。本研究将利用 SEER 数据库, 提取其中的亚洲人 NSCLC 病例, 采用更能反映预后变量之间相关关系且适用性更好的机器学习方法, 构建亚洲人 NSCLC 预后模型及预测评估体系, 为临床医生开展

通讯作者: 赵玉虹, ORCID: 0000-0003-4265-6692, E-mail: joan@mail.cmu.edu.cn。

*本文系国家自然科学基金项目“中国临床医师岗位胜任力模型构建及评价体系研究”(项目编号: 71473268)、辽宁省科学技术计划项目“肝炎、结核等重大疾病临床研究平台建设”之子项目“构建辽宁(本溪)生物医药科技产业基地的信息化服务与成果转化创新平台”(项目编号: 2013225079)和教育部人文社会科学研究青年基金项目“基于语义述谓网络属性的多文档自动摘要: 以生物医学为例”(项目编号: 13YJC870030)的研究成果之一。

治疗与判断预后提供决策支持。

2 相关研究

国内外对疾病预测模型的研究已经有一定基础。Muers 等[3]获取 6 所医疗机构中 NSCLC 患者数据建立其预后风险模型，并将模型所输出的生存期与临床医生的判断作比较; Yang 等[4]基于 SEER 数据库构建横纹肌肉瘤患者 5 年及 10 年生存预测模型以指导治疗方法的选择; Park 等[5]利用临床试验数据，预测采用姑息化疗的晚期胆道腺癌患者的生存情况。以上模型均先筛选预后因素再采用统计学中的 COX 回归方法构建模型，这也是构建医学预测模型的常见方法。但 COX 回归分析难以看出预后变量之间的关系，为提高模型的适用性，机器学习方法逐渐受到研究者的推崇。如 Kim 等[6]应用朴素贝叶斯方法并绘制诺模图(Nomogram)，通过 7 项指标得到术后复发的可能性，该研究者曾于 2012 年使用支持向量机方法预测乳腺癌患者术后 5 年生存情况[7]，而后构建在线预后系统。

自 21 世纪初开始，国内越来越多的研究者开始从机器学习方向出发评价肿瘤及其他疾病的发生、发展和预后。刘雅琴[8]基于 SEER 数据库使用 Logistic 回归、人工神经网络、决策树三种方法比较预后预测模型效果，是国内此领域研究肿瘤预后的重要突破。台湾学者 Chen 等[9]使用人工神经网络对 4 个医疗机构的 NSCLC 患者的临床及基因表达数据进行探究，建立生存状况风险模型; 牟冬梅等[10]通过提取电子病历信息进而构建妊娠高血压综合征危险因素预测模型，得到决策树模型为最优。但是以上研究的变量纳入均凭借已有经验，缺少与临床医生的交流，未实现跨学科的合作。

通过文献研究发现: 肿瘤中发病率及死亡率均较高的肺癌的预后研究屈指可数。因此，本文基于 SEER 数据库，确定患者预后因素并参考肿瘤医生的意见进行调试，利用更能反映预后变量之间相关关系且适用性更好的机器学习方法，以提升预测准确率为目标，构建亚洲 NSCLC 患者术后生存模型，更好地为临床预后评价服务。

3 肿瘤预后模型构建方案

肿瘤的预后包括风险评估、复发、转移及生存情况评价[11]。以 NSCLC 患者术后 5 年为时间基准，对患者的生存情况即“生存”与“死亡”进行预测，具体研究流程如图 1 所示。

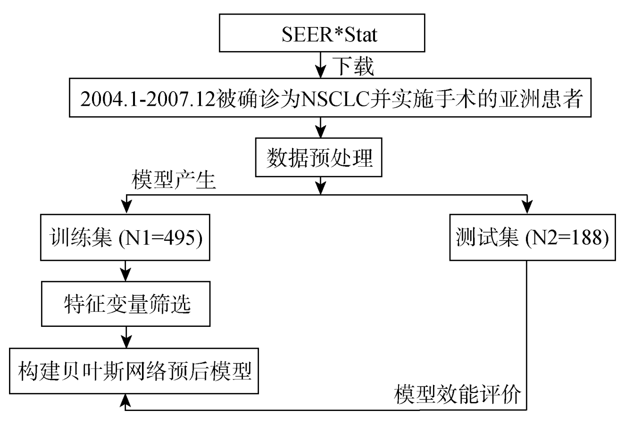


图 1 基于 SEER 构建亚洲 NSCLC 患者预后模型的研究流程

具体步骤如下:

(1) 数据下载: 在 SEER*Stat 软件中调用 Incidence-SEER18 Regs Research Data+Hurricane Katrina Impacted Louisiana Cases, Nov2014 版本数据，该版本数据随访终止日期为 2012 年年末，并根据 ICD-O-3 恶性肿瘤形态学编码，下载 NSCLC 患者数据。

(2) 变量选取依据: 参考美国癌症联合会(American Joint Committee on Cancer,AJCC)、美国国立癌症网络(The National Comprehensive Cancer Network, NCCN)临床指南及美国第二版肿瘤信息采集系统[12-13](Collaborative Stage Manual Online Help, CS)中所提及与患者生存相关的预后因素，并从 SEER*Stat 中提取含有上述变量的所有字段，以首次确诊时所登记的患者信息为准，将整理后的患者数据录入 Excel 表。

(3) 特征变量筛选: 为确定各变量是否独立影响患者的生存情况，首先应用 SPSS22.0 软件对训练样本进行单因素分析(独立样本 t 检验或卡方检验)，而后将经单因素分析得到的变量纳入 Logistic 回归分析，并筛选 NSCLC 高相关预后因素，$P<0.05$ 具有统计学意义，结合临床医生的建议调整变量纳入最终模型。

(4) 肿瘤预后模型的构建: 选用机器学习中的监督学习方法，进行肿瘤预后预测模型的构建[10]。应用 R Studio 软件建立贝叶斯生存预测模型，并完成贝叶斯网络的结构调整，构建有效的预后模型。

(5) 模型评价: 选用数据挖掘软件 WEKA 比较贝

叶斯网络模型及其他三种常见分类模型的预测准确性、精确度及 ROC 曲线下面积。

4　研究过程

4.1　肿瘤预后模型的构建

(1) 研究对象

选取自 2004 年起被确诊为 NSCLC 的亚裔患者为最终研究对象，其中包含 5 年内直接因 NSCLC 致死和随访期满 5 年且仍然生存的患者，共计 683 位。

(2) 研究变量

在 SEER 中提取 17 项预后研究变量：性别、国别、婚姻状况、发病部位、病理类型、组织学分级、患侧部位、邻近器官浸润程度、区域淋巴结累积程度、远处转移程度、肿瘤分期、手术类型、是否接受放疗以及确诊时年龄、肿瘤大小、阳性淋巴结数量及受检淋巴结数量，其中后 4 项指标为连续型变量，其余均为分类变量，如表 1 所示。

表 1　非小细胞肺癌患者预后指标信息

数据类型	变量	SEER 中所示名称	类数/数值范围
分类型	性别	Sex	2
	国别	Race recode (Asian)	8
	婚姻状况	Marital status at diagnosis	4
	发病部位	Primary Site - labeled	5
	病理类型	ICD-O-3 Hist/behav, malignant	4
	组织学分级	Grade	4
	患侧部位	Laterality	2
	邻近器官浸润程度	CS extension	18
	区域淋巴结累积程度	CS lymph nodes	5
	远处转移程度	CS mets at dx	5
	肿瘤分期	Derived AJCC Stage Group	7
	手术类型	RX Summ--Surg Prim Site	13
	是否放疗	Radiation	3
连续型	确诊时年龄	Age at diagnosis	26-90
	肿瘤大小	CS tumor size	4-132
	阳性淋巴结数量	Regional nodes positive	0-23
	受检淋巴结数量	Regional nodes examined	1-45

(3) 结局变量

肿瘤患者 5 年生存情况是评价预后效果的重要指标。以 NSCLC 患者术后 5 年的生存情况作为应变量。其中生存期以月为单位，对其进行分类变量的转换，即生存时间在 60 个月及以上的患者被视为“生存”(记为 1)，否则即为“死亡”(记为 0)。

(4) 特征变量选择

为减少预后变量，提高模型的预测准确性，需对纳入研究变量进行高相关预后因素选择。经单因素分析后初步纳入的变量有($P<0.05$)：确诊时年龄、肿瘤大小、组织学分级、肿瘤分期、邻近器官浸润程度、区域淋巴结累积程度、阳性淋巴结数量、婚姻状况、国别、远处转移程度、手术类型及是否放疗。在单因素分析的基础上经 Logistic 回归分析筛选出的预后变量有($P<0.05$)：确诊时年龄、肿瘤大小、组织学分级、肿瘤分期、受检淋巴结数量及阳性淋巴结数量。筛选结果如表 2 所示。

表 2　Logistic 回归分析筛选变量结果

变量名称	B	S.E.	Exp(B)	95% Exp(B)		Sig.
				下限	上限	
确诊时年龄	−0.066	0.011	0.936	0.916	0.957	0.000
肿瘤大小	−0.018	0.007	0.982	0.968	0.996	0.014
组织学分级	/	/	/	/	/	0.001
肿瘤分期	/	/	/	/	/	0.013
受检淋巴结数量	0.050	0.017	1.051	1.016	1.087	0.004
阳性淋巴结数量	−0.199	0.067	0.819	0.719	0.934	0.003

受累淋巴比率(Lymph Nodes Ratio, LNR)为阳性淋巴结数量与受检淋巴结数量的比值，参考临床医生的意见，将 LNR 代替阳性淋巴结数量和受检淋巴结数量两项作为预后变量，即最终进入模型的变量为：确诊时年龄、肿瘤大小、组织学分级、肿瘤分期及受累淋巴比率。

(5) 数据预处理

删除数据缺失严重、记录错误及因非肺癌致死的患者信息，选用 Interval 方法对数值型数据进行离散化。该离散化方法旨在将区间 $[X_0, X_{N-1}]$ 划分为同样大小的子区间 D 并根据所属子区间指数给出离散化意见，其中观察指数 i 与离散水平 j 满足以下条件[14]:

$$X_0+\frac{j(X_{N-1}-X_0)}{D}<X_i\leqslant X_0+\frac{(j+1)(X_{N-1}-X_0)}{D}$$

在软件R Studio中调用bnlearn的函数包实现以上数据预处理步骤。而后按照约 70%与 30%的比例[15]将数据分为训练集(N_1=495)和测试集(N_2=188)，训练集用来进行网络学习及调整，从而构建预后模型，测试集则用来评价模型的性能。

(6) 预后模型的构建及预测结果

贝叶斯网络(Bayesian Network, BN)通过表示变量的节点和表示变量间关系的连线描述子节点与父节点间的依赖关系[16]，已知随机变量 $X=\{X_1, X_2, \cdots, X_n\}$，其联合概率分布为:

$$P(X_1,\cdots,X_n)=\prod_{i=1}^{n}P(X_i\mid Pa(X_i))$$

其中，$Pa(X_i)$是X_i父节点的子集，在网络图中X_i独立于其非直系节点变量。选用禁忌搜索(Tabu Search, TS)方法对贝叶斯网络进行初步学习。该方法于 1986 年由美国工程学院院士Fred Glover提出[17]，是一种基于邻域和迭代来求解优化问题的启发式算法。该方法的本质是禁止重复前面的工作，跳出局部搜索最优点，即在区域中随机移动并产生新的方案，而后将评估每一个相邻的解决方案，并选择最能提高目标函数的路径，若没有能提高最终结果的方案，则选取对目标函数影响最小的方案，通过模仿人类记忆找出最佳结果[18]，步骤如下:

①确定区域 $N(x)$，从中选定一个初始可行解 X^0，使当前最优解 $X^{best}=X^0$，则 $T=N(X^{best})$;

②按照上述步骤依次组合，并得到最新解 X^{next_n}，$n\in[1,+\infty]$，输出计算结果;

③比较所有决策结果并输出全局决策最优解 $X^{next_{max}}=X^{best}$。

Makond 等[19]所构建的贝叶斯预后模型并未完全根据所得数据进行学习，而是通过听取医生意见建立患者预后生存模型，实际上是基于实际经验的建模思维。本研究克服单以实际经验建模的弊端，结合网络学习方法TS与医生意见共同建立患者预后模型，在R Studio 软件中实现网络模型的修整与优化，最终的网络模型如图2所示。

在R Studio软件中使用caret包输出预测样本及实例所组成的表格及预测模型评价指标。本研究共188个测试集样本，预测正确137例，预测正确率达72.87%。

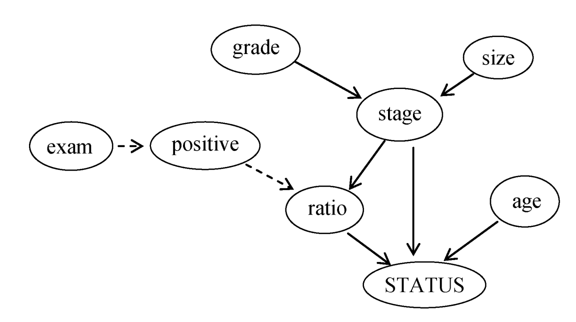


图 2　亚洲非小细胞肺癌患者预后生存贝叶斯网络模型

4.2　对比实验

另选用决策树、支持向量机及人工神经网络方法建立预后模型，并根据预测结果与本研究所构建的预后模型作对比。在WEKA中分别选择三种方法所对应的J48、SMO及Multilayer Perceptron建立预后模型，参数默认。4 种机器学习算法建模的预测准确性及模型性能评价比较如表4–表5所示。

表 4　BNNSCLC 模型与其他三种分类算法所建模型预测准确率比较

所用分类算法	预测准确率	
	训练集	测试集
贝叶斯网络	0.683	0.729
决策树	0.713	0.670
支持向量机	0.733	0.686
人工神经网络	0.784	0.649

表 5　不同算法所构建模型性能比较

算法	预测准确率	精确度	ROC 曲线下面积
贝叶斯网络	72.87%	71.0%	0.67
决策树	67.02%	66.3%	0.568
支持向量机	68.62%	68.2%	0.611
人工神经网络	64.89%	63.7%	0.615

4.3　实验分析

本研究发现贝叶斯网络所构建的 NSCLC 预后模型最优。由表 4 可知，虽然决策树、支持向量机及人工神经网络在训练集上的预测准确性均高于贝叶斯网络，但在测试集中三者预测准确性的数值与训练集相比显著下降，未能很好地适应新数据，不适于实际应用，模型的拟合程度不如贝叶斯网络模型。另通过对表 5 的解读，贝叶斯网络模型在预测准确率、精确度及 ROC 曲线下面积的数值均高于其他三种机器学习算法。

网络学习方法的选择是构建贝叶斯分类器的基础。本研究选用 TS 方法初步对网络模型进行构建，是对爬山法的优化，当已知构成某网络变量并不产生网络环路的基础上，以移动搜索代替随机产生，采用加、减及逆向边三种操作产生邻域[20]，并搜索全局最优解来调整网络结构以完成贝叶斯网络的自学习。在此基础上，结合临床医生的经验修改网络图，将高相关预后因素相联系，是理论方法与实际应用的典型结合。

网络图的调整是该生存预测模型构建研究的最关键流程。如图 2 所示，箭头方向表示节点间的关系，如 size 指向 stage 即为前者直接对后者产生影响，所选预后变量均指向最终变量生存状态，其中确诊时年龄、肿瘤分期及受累淋巴比率直接影响患者的生存情况。通过构建不同的网络图找到最优分类模型，从而判断各预后因素间的关系及对生存状态的影响，临床可据此评价肿瘤患者术后的预后情况，并对相关因素进行控制。当然，由于本研究所采用的 SEER 数据库并未将所有肿瘤预后因素全部纳入库中[21]，故建模所选指标的数量有限，该预测模型可能存在一定局限性。

5　结　语

本研究以非小细胞肺癌患者术后生存状态为目标构建患者生存预后模型，预测准确率达 72.87%。通过构建贝叶斯网络探寻预后变量间的关系及对患者生存情况的影响，在网络结构内部调整的基础上结合临床专家的建议，更好地诠释了模型中节点间关系。首次应用 SEER 数据库，以亚洲肿瘤患者为主要研究对象构建其生存预测模型，对判断患者术后 5 年的预后情况起到辅助作用，具有应用前景。在未来的研究中，可考虑其他来源患者外部验证的纳入，提升预测模型自身的适应程度，更好地为临床治疗及预后评价服务。

参考文献：

[1] National Cancer Institute. SEER Cancer Statistics Review (CSR) 1975-2013 [R/OL]. [2016-09-20]. http://seer.cancer.gov/csr/1975_2013/sections.html.

[2] Ettinger D S, Wood D E, Akerley W, et al. NCCN Guidelines Insights: Non-Small Cell Lung Cancer, Version 4.2016 [J]. Journal of the National Comprehensive Cancer Network: JNCCN, 2016, 14(3): 255-264.

[3] Muers M F, Shevlin P, Brown J. Prognosis in Lung Cancer: Physicians' Opinions Compared with Outcome and a Predictive Model[J]. Thorax, 1996, 51(9): 894-902.

[4] Yang L, Takimoto T, Fujimoto J. Prognostic Model for Predicting Overall Survival in Children and Adolescents with Rhabdomyosarcoma[J]. BMC Cancer, 2014, 14: 654. DOI: 10.1186/1471-2407-14-654.

[5] Park I, Lee J L, Ryu M H, et al. Prognostic Factors and Predictive Model in Patients with Advanced Biliary Tract Adenocarcinoma Receiving First-line Palliative Chemotherapy [J]. Cancer, 2009, 115(18): 4148-4155.

[6] Kim W, Kim K S, Park R W. Nomogram of Naive Bayesian Model for Recurrence Prediction of Breast Cancer [J]. Healthcare Informatics Research, 2016, 22(2): 89-94.

[7] Kim W, Kim K S, Lee J E, et al. Development of Novel Breast Cancer Recurrence Prediction Model Using Support Vector Machine [J]. Journal of Breast Cancer, 2012, 15(2): 230-238.

[8] 刘雅琴. 乳腺癌患者预后模型的研究[D]. 上海: 上海交通大学, 2008. (Liu Yaqin.Study on the Prognosis Model for Breast Cancer [D]. Shanghai: Shanghai Jiaotong University, 2008.)

[9] Chen Y C, Ke W C, Chiu H W. Risk Classification of Cancer Survival Using ANN with Gene Expression Data from Multiple Laboratories [J]. Computers in Biology and Medicine, 2014, 48: 1-7.

[10] 牟冬梅, 任珂. 三种数据挖掘算法在电子病历知识发现中的比较[J]. 现代图书情报技术, 2016(6): 102-109. (Mu Dongmei, Ren Ke. Discovering Knowledge from Electronic Medical Records with Three Data Mining Algorithms [J]. New Technology of Library and Information Service, 2016(6): 102-109.)

[11] Shin H, Nam Y. A Coupling Approach of a Predictor and a Descriptor for Breast Cancer Prognosis [J]. BMC Medical Genomics, 2014, 7(S1): S4.

[12] American Joint Committee on Cancer, AJCC Cancer Staging Manual [M]. The 7th Edition. New York: Springer Verlag, 2010: 253-270.

[13] National Comprehensive Cancer Network: NCCN Clinical Practice Guidelines in Oncology: Non-Small Cell Lung Cancer, Version 2.2016 [R/OL]. [2016-09-20]. http://www.nccn.org/patients.

[14] Hartemink A J. Principled Computational Methods for the Validation and Discovery of Genetic Regulatory Networks [D]. Massachusetts Institute of Technology, 2001: 86-87.

[15] Kumar Y, Sahoo G. Prediction of Different Types of Liver Diseases Using Rule Based Classification Model [J]. Technology & Health Care Official Journal of the European Society for Engineering & Medicine, 2013, 21(5): 417-432.

[16] Oh J H, Craft J, Al L R, et al. A Bayesian Network Approach for Modeling Local Failure in Lung Cancer [J]. Physics in Medicine & Biology, 2011, 56(6): 1635-1651.

[17] 张雪雷. 基于禁忌搜索算法的贝叶斯网络在疾病预测与诊断中的应用[D]. 太原: 山西医科大学, 2015. (Zhang Xuelei. The Application of Bayesian Network Based on Tabu Search Algorithm in Diseases Prediction and Diagnosis [D]. Taiyuan: Shanxi Medical University, 2015.)

[18] Lim W L, Wibowo A, Desa M I, et al. A Biogeography-Based Optimization Algorithm Hybridized with Tabu Search for the Quadratic Assignment Problem [J]. Computational Intelligence & Neuroscience, 2016. DOI: 10.1155/2016/5803893.

[19] Makond B, Wang K J, Wang K M. Probabilistic Modeling of Short Survivability in Patients with Brain Metastasis from Lung Cancer [J]. Computer Methods & Programs in Biomedicine, 2015, 119(3): 142-162.

[20] 魏珍, 张雪雷, 饶华祥, 等.禁忌搜索算法的贝叶斯网络模型在冠心病影响因素分析中的应用[J].中华流行病学杂志, 2016, 37(6): 895-899. (Wei Zhen, Zhang Xuelei, Rao Huaxiang, et al. Using the Tabu-search-algorithm-based Bayesian Network to Analyze the Risk Factors of Coronary Heart Diseases [J]. Chinese Journal of Epidemiology, 2016, 37(6): 895-899.)

[21] 杨乔, 张俊萍. 肿瘤登记数据库的临床应用[J]. 循证医学, 2013, 13(4): 250-251, 256. (Yang Qiao, Zhang Junping. Clinical Applications of the Tumor Registry Database [J]. The Journal of Evidence-Based Medicine, 2013, 13(4): 250-251, 256.)

作者贡献声明:

尹玢璨: 设计研究方案, 数据分析, 构建模型, 撰写论文;
辛世超: 数据预处理, 建模实验;
张晗: 修改论文;
赵玉虹: 提出研究思路, 论文最终版本修订。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: yinbincan0803@163.com。

[1] 尹玢璨. NSCLC.csv. 亚洲非小细胞肺癌患者预后模型研究原始数据.

[2] 尹玢璨. data.csv. 亚洲非小细胞肺癌患者建模数据.

收稿日期: 2016-10-31
收修改稿日期: 2016-12-05

Building Asian Tumor-patients Prognostic Model with Bayesian Network and SEER Database——Case Study of Non-Small Cell Lung Cancer

Yin Bincan[1] Xin Shichao[1] Zhang Han[1] Zhao Yuhong[1, 2]
[1](Department of Medical Informatics, China Medical University, Shenyang 110122, China)
[2](Shengjing Hospital of China Medical University, Shenyang 110004, China)

Abstract: [**Objective**] This study aims to improve the tumor-prognostic assessment for Asian patients who were diagnosed with Non-Small Cell Lung Cancer (NSCLC). The proposed model identifies the influencing factors of the patients' survival status and predicts their prognostic situation. [**Methods**] First, we used single factor statistical method and logistic regression to identify the prognostic variables. Second, we employed the Bayesian Network algorithm to construct the prognostic survival model for the Asian NSCLC patients. Finally, we compared the performance of our model with three other algorithms. [**Results**] The identified prognostic variables include age, tumor size, grade, tumor stage, as well as the lymph nodes ratio. The proposed model could predict NSCLC patients' prognostic survival status effectively. [**Limitations**] The SEER database had limited number of prognostic factors, which may influence the prediction accuracy. [**Conclusions**] The Bayesian Network could help us build optimal prognosis model for cancer patients to improve their survival rates. The proposed model is better than the Decision Tree, Support Vector Machine and Artificial Neural Network models.

Keywords: Bayesian Networks Non-Small Cell Lung Cancer Prognosis Machine Learning